智库信息组织策略及其在大数据环境下的挑战*

[1,2]安楠 [1]祝忠明
1 中国科学院兰州文献情报中心 兰州 730000
2 中国科学院大学 北京 100049

摘要：[目的/意义]高水平的新型智库离不开高水平的信息支持机制，大数据时代背景下传统的智库信息组织机制已无法适应当前的数据特征及决策要求，构建支持决策过程的知识库已成为智库发展的必然趋势。[方法/过程]本文选取《全球智库报告 2016》中具有参考价值的国外智库机构作为研究对象，应用文献调研法和案例分析法总结归纳了目前智库常见的信息组织方式，分析了大数据下数据价值链及其对组织环节的要求，并据此提出智库知识库构建的必要性。[结果/结论]最终提出一个通用的面向决策过程的智库知识库框架，并采用语义本体方法构建了知识库内部的知识组织模型，以期为智库在大数据下逐渐实现半自动到自动化的决策研究过程提供参考借鉴。

关键词：智库;知识库;大数据;信息组织;组织策略;决策研究;本体
分类号：G359

1 引言

智库是公共政策的研究分析和参与机构，它们针对国内、国际问题开展政策导向性的研究、分析和咨询，以使得政策制定者和公众能够依据可靠的信息进行决策[1]。其主要作用是为决策制定者提供及时、全面、准确的支持信息，支持信息的范围、数量、质量、服务内容、服务方式等都将直接影响到决策制定的效果[2]，因此拥有完善的信息支持机制是智库产生高质量决策咨询成果的重要保障。

大数据时代信息呈现出体量巨大、形式繁多、更新速度快及价值密度低的4V 数据特征[3][4]，在这种数据爆炸的形势下，任何研究过程都呈现出一种数据驱动的趋势，如何从海量信息中及时发现、提取有价值的知识为自己所用，将成为影响智库决策研究过程及产出效率的关键。

大数据驱动下的智库决策研究必须解决以下两个问题：一是如何构建一个统一的数据模型，使得任何大数据资源都能够通过该数据模型的加工处理最终成为可支持决策研究的智能数据，逐渐实现半自动到自动化的决策研究过程。二是如何针对决策研究过程对各种来源各种形式的相关信息进行语义化处理，加强数据之间的关联以提升知识发现的能力，为决策者提供更有价值的政策参考信息。因此，本文将围绕大数据分析能力需求及智库决策研究过程尝试构建支持多源异构的数据集成框架，为语义化地建造支持大数据情报处理和分析的智能数据集提供统一的概念模型。

2 智库信息组织机制现状研究

智库如何对数据内容进行组织加工将直接影响到研究人员与情报专家对信

息资源的利用效率，科学合理的组织方式不仅能提高数据存取效率，更有助于挖掘数据中的潜在价值信息，产生增值效应。

2.1 国外智库信息组织机制发展现状

现代意义上的智库最早形成于二战时期的西方国家，相比我国，西方国家无论在智库研究领域还是智库自身建设都已发展的相对完善，选取西方有影响力的智库作为研究对象将更具代表性。本文依据宾大《全球智库报告 2016》的综合排名及各项领域排名，选取了排名靠前的十余家具有代表性的国外智库作为研究对象（见表 1），通过对其官方网站上信息资源的展示方式以及可获取的各种类型智库产品的调研，对其信息组织策略进行了分析。

表 1 调研涉及的国外智库

Table 1 Foreign Think Tanks involved in the research

智库名称	所属国家
布勒哲尔国际经济研究所（Bruegel）	比利时
斯德哥尔摩国际和平研究所（SIPRI）	瑞典
世界资源研究所（WRI）	美国
卡内基国际和平基金会（CEIP）	美国
兰德公司（RAND）	美国
美国中央情报局（CIA）	美国
德国国际与安全事务研究所（SWP）	德国
胡佛研究所（HOOVER Institution）	美国
布鲁金斯学会（Brookings Institution）	美国
卡托研究所（Cato Institute）	美国
查塔姆研究所（Chatham House）	英国
马普学会（Max Planck Society）	德国
日本国际问题研究所（JIIA）	日本

通过调研比较与分析，归纳了西方智库机构常见的信息搜集及组织策略（见图 1），总结了当前西方智库信息支持机制的发展现状。在智库的信息搜集策略中，主要以需要较多依靠人工参与的手动采集和半自动采集为主，其中搜集公开数据以其可操作性较强、数据范围广、相对成本低等特点成为智库最常用的信息搜集方式之一，几乎所有上述智库都将通过互联网获取公开数据作为数据搜集的最常规途径。此外因智库研究的实时性和新颖性，智库经常对所需数据有特殊要求或涉及到诸如战争形势、行为科学、药物病理等特定项目，没有完全适用的数据或先前数据参考价值不大，因此智库研究人员还需通过直接生产创造途径作为对间接搜集获取途径的补充，其中文献调查法因其低成本且易开展成为使用频率最高的直接获取数据方式，例如美国布鲁金斯学会、胡佛研究所、卡内基国际和平基金会等老牌智库在传统调研运用中都是最典型的代表。当然，在调查研究过程中智库专家经常不拘泥于某种特定方法，而是相互交错、灵活运用。依据内容的组织形式，搜集到的信息资源可被组织为数据库、信息检索系统、知识库三种

形式（详见图 1）。

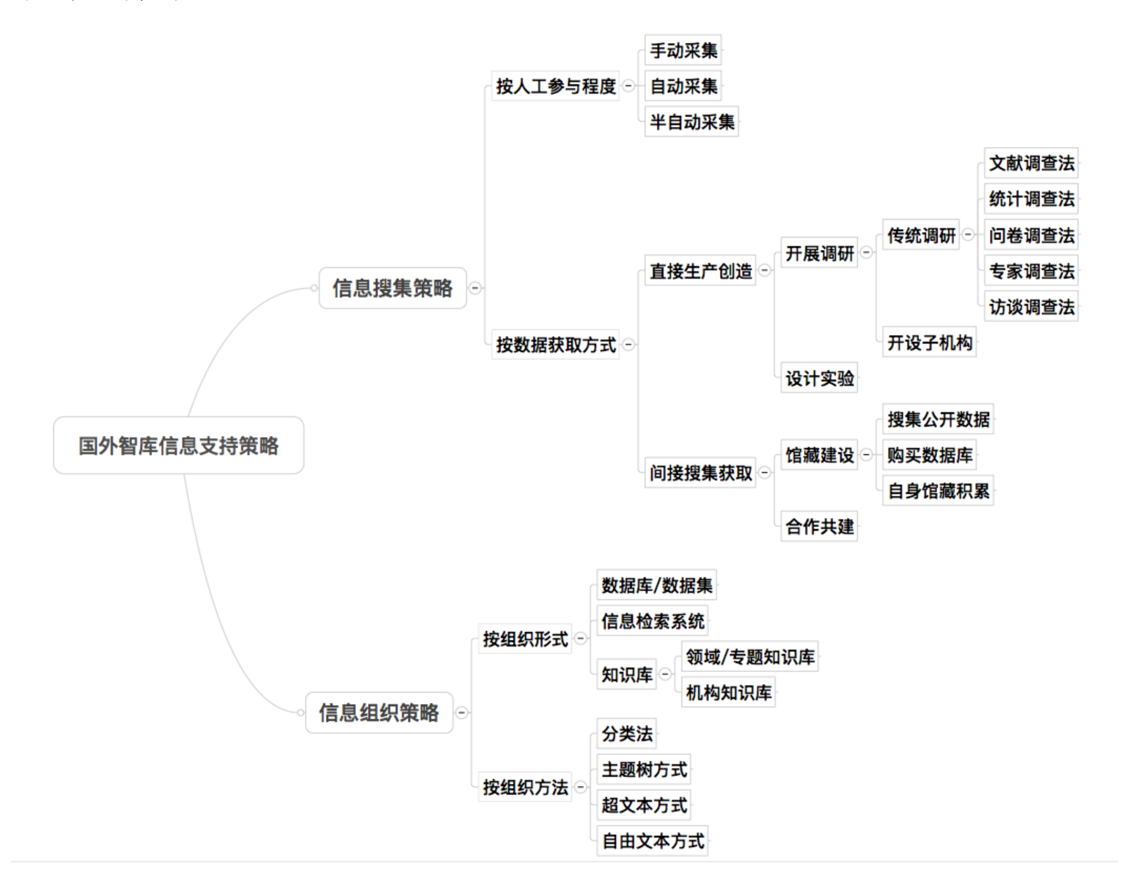


图 1 国外智库信息支持策略

Figure 1 Information support strategies of foreign think tanks

2.1.1 数据库（数据集）

对于智库通过直接或间接途径搜集到的数据，组织方式之一就是将其结构化为数据库或数据集，这种结构化数据形式的优点是便于管理、共享性高、冗余度低、容易扩充。

布勒哲尔国际经济研究所（Bruegel）是一家专注于国际经济政策研究的智库， 其将关于政策经济的 7 个专业数据集对外开放[5]，包括（1）全球及地区基尼系数；（2）欧元区货币总量 Divisia 指数；（3）178 个国家的实际有效汇率；（4）全球经济下的欧洲企业：外部竞争下的内部政策；（5）持有的主权债券；（6）欧元体系流动性；（7）在 PATSTAT 应用程序上基于回归的记录链接。斯德哥尔摩国际和平研究所（SIPRI）以其对全球安全问题权威性的评估享誉世界，SIPRI 所有研究的根据和来源均完全开放，因此其研究成果成为国际政治家、研究人员及媒体人员经常使用的权威性资料来源。SIPRI 拥有 4 个专业数据库：（1）多边和平行动数据库；（2）军费开支数据库；（3）武器转让数据库；（4）军需工业数据库。此外，SIPRI 还全面掌握了关于军备控制和裁军的数据集[6]，包括军火禁运报告、国家军火报告、全球军火贸易价值报告等等，这些专业数据库对 SIPRI 的研究活动提供了强有力的信息支持。

2.1.2 信息检索系统

一个机构信息检索系统的完善程度也可以直接反映出其信息组织的好坏，对于智库来说，强大的检索系统不仅能从内部为研究专家提供高效率的数据支持，同时为用户快速准确地获取所需信息提供了便利。

卡内基国际和平基金会提供了简洁易用的站内检索系统，用户可选择精确匹配或任意匹配的方式对题名、作者名或全文进行检索，检索结果可通过文档类型、发表年份、地区、主题、项目进一步筛选，并可按照日期或相关度对结果进行排序。兰德公司的检索系统功能相对完善，用户可以通过关键词匹配、额外属性、文档特征等多种检索条件进行限定，额外属性包括页面标题、所属兰德部门、内容类型、起始日期等，文档特征涵盖了题名、作者、主题、ISBN 等用以快速定位到相关资源。美国中央情报局（CIA）作为美国乃至世界著名的情报机构之一，力求对海量情报进行科学管理使其效果得到最大程度发挥，这也令 CIA 成为情报机构中进行信息资源管理与增值的典范。CIA 的解密档案检索系统在对档案材料进行数字化保存时采用了元数据方法（见表 2），统一的元数据标准将海量信息资源进行科学归类，同时能够将文本、音视频等不同类型的媒介资源进行有机融合，使其在同一个存取体系内进行统一检索，极大提高了信息利用效率。

表 2 CIA 解密文档检索系统元数据

Table 2 Metadata of CIA decrypted document retrieval system

CIA 解密档案检索系统中使用的 10 种元数据	
文件类型 Document Type	专藏 Collection
文件编号 Document Number	公开决定 Release Decision
文件页数 Document Page Count	原始密级 Original Classification
文件附件 File Attachment	序列号 Sequence Number
案件编号 Case Number	出版日期 Publication Date

2.1.3 知识库

在信息环境中知识库（knowledge repository）可以被定义为一个组织围绕特定应用目的（如支持科研、教育或管理过程等）建立的知识集合。一般地，知识库有两种基本的类型：领域/专题知识库和机构知识库。前者收集、组织和传播特定学科领域或主题的知识内容，后者主要提供对一个机构产出的知识进行保存和传播管理的服务。知识库作为一种存储、组织和管理数字知识的机制，在科研领域已经有着较为广泛的应用，然而在智库等决策咨询机构中的应用还尚不成熟，相当一部分智库由于资金、资源等原因或者还没有意识构建智库内部的知识库，仍停留在信息“存储库”的阶段。

本文结合之前已有的研究[7]，通过调研从馆藏建设、情报搜集、技术支持三个方面对比分析了美国兰德公司（RAND）和德国国际和安全事务研究所（SWP）

在知识库建设过程中的情况（见表 3），通过分析可以看出，RAND 和 SWP 都非常重视对信息资源的建设，内部馆藏丰富，数据库内容涉及广泛。在知识组织方面均采用了分类组织的方式，依据研究主题建立专题知识库，同时也选择地区作为研究项目分类的依据。两大智库均通过技术手段开发了信息支持系统，并且都积极尝试与其他机构部门开展信息资源共享、信息共建等合作，以弥补自身专业缺陷，同时能减少数据冗余。

表 3 RAND 与 SWP 知识库构建情况对比

Table 3 Comparison of knowledge repositories between RAND and SWP

	馆藏建设			情报搜集		技术支持	
	内部图书馆	数据库	项目报告	人员部门	子机构	开发系统	合作
RAND	55000 本图书、134000 份报告、3000 种期刊、4000 张地图，以及特殊形式的文件和缩微品。	涉及内容包括健康、犯罪、安全、灾害、军事、网络、金融、社会、医疗、教育、能源、人口调查、劳动力、就业、收入等主题。	研究项目被按照地区、主题划分，另有政策聚焦、热点趋势、特色专题研究活动等板块。	于 1972 年成立兰德调查组 SRG，在全球范围内进行调查数据的搜集，通过收集到的数据为用户提供数据分析与规划服务；聘用 900 多位知名教授及各领域专家作为特约顾问和研究员。	设有兰德欧洲、澳洲分部，以及亚太政策中心、中东政策中心、俄罗斯及欧亚中心、全球安全中心。	基于 RITA 语言设计了知识库专家系统，帮助研究人员分析恐怖分子活动；开发了 RaDiUS，提供深层次的信息资源共享。	与美国联邦政府各部门开展联合研究；与多国图书馆建立馆际互借关系，满足跨地区跨领域的信息需求。
SWP	92000 册藏书、440 种杂志、380 部年鉴、130 种报纸。	与德国国际事务和地区研究信息网络（FIV）协作，FIV 是基于 12 个德国专题研究院构建的统一结构化、集成化的公开访问数据库。	研究项目被按照地区、主题划分，另有已完成项目的查看入口。	5 个按地区划分的研究部门以及国际安全与全球事务 2 个部门；拥有研究人员 73 名，另专设 38 名情报与知识管理人员负责信息管理、图书馆、信息网络与系统管理、信息分配等工作。	除了位于德国柏林的总部，于 2009 年在比利时布鲁塞尔开设办公室，保障了与北约和欧盟的活跃交流，并与欧洲各研究所和智库保持联系。	开发了国际关系与地区研究文献检索系统 IREON，提供包括 WAO、PAIS 等科学文献搜索服务和全文链接。	与欧洲其他研究机构合作开发了针对国际关系和地区研究的词库 European Thesaurus，考虑到用户的全球性，该词库支持德、英、法、意、俄等 9 种语言。

2.2 大数据时代对智库信息组织的挑战

由调研可以看出，虽然当下全球各大智库的信息支持机制已发展得较为全面，但是仍存在很多不足。一方面，以数据库/数据集形式组织起来的信息之间是相对独立的，即使数据库对所存储的信息有从主题或其他特征进行大致分类，但在更细粒度层面的数据上，数据之间彼此独立，缺乏必要的关联，不利于智库在进行数据挖掘和数据分析时对潜在知识的发现。此外，这种数据相互独立的信息组织方式没有基于上下文（context-based）的联系，缺乏语境化和情景化的知识应用，即针对同一概念在不同情景下的理解能力较弱。另一方面，在以信息检索系统形式组织信息的智库中，绝大部分仅仅标注了信息的外延，并没有针对信息内容进行更深层的语义化标注，不利于计算机对数据信息的理解以及智库决策过程自动化的发展，这在大数据时代智库对国际形势响应速度要求越来越高的情况

下显然已经阻碍到了智库的决策产出效率。为了能够将大数据中的无意义数据加工为可支持决策研究的智能数据，各种信息必须从非结构化的、彼此独立存在的粗粒度数据被加工成结构化的、计算机可操作的、相互关联的、具有上下文语境的细粒度数据。

全球知名咨询公司麦肯锡最早提出了“大数据”时代的到来，大数据时代各种数字资源急剧增长，逐渐成为信息资源的主流。面对大数据的 4V 特征，传统的智库信息支持机制已无法高效处理如此海量的异构数据，如何有效地从纷杂的数据中获取有价值的信息，如何对采集到的海量信息进行科学的管理和组织，并以此为用户提供迅速、准确的服务，这就要求新型智库必须及时调整对数据的采集、存储及组织策略以适应当前的大数据特征。 T.Gustafson 和 D.Fink 于 2013 年提出“大数据价值链”的概念[8]，认为每条大数据价值链简化后都至少应由 4 个基本阶段组成：数据获取——数据存储——数据分析——数据应用。智库作为知识组织型机构，其决策研究及决策产出过程实际上也是一个知识增值的过程。基于此，提出大数据环境下的智库数据价值链（如图 2），智库数据价值链反映了智库决策研究及产出的各个阶段围绕数据进行的活动，而大数据则为各环节提出了要求。

与一般依赖计算机自动化抽取、处理并分析大数据得到结果的商业化研究的数据价值链不同，智库数据价值链进行知识增值的过程是一个基于前者对大数据的充分处理和组织后，作为决策研究的支持数据提供给智库专家，与智库专家的隐形知识共同作用最后形成智库产品的过程。在这种数据驱动的决策过程下，最终提供给智库专家的支撑数据的及时性、全面性、准确性以及数据组织完善程度都将直接影响到最后智库产品的产出效率和质量。

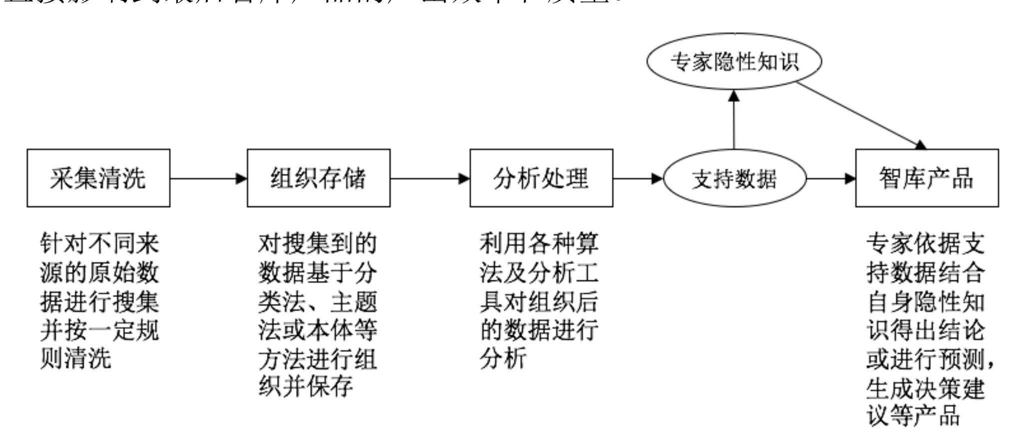


图 2 大数据环境下的智库数据价值链

Figure 2 The data value chain of think tanks in the context of big data

从智库数据价值链中对各环节的要求可以看出，大数据时代对传统智库研究的一系列流程都产生了影响，其中最关键的应是处于中间环节的数据组织阶段。作为承前启后的中间环节，智库在开展数据组织工作时既要适应之前智库从各种

数据源采集的复杂数据类型，又需为后续进行存储及数据分析要使用的技术和工具提前做好相应准备。综上，大数据下智库需要一个能够结合管理手段和信息技术对捕获并保存到的信息进行有效组织和管理的信息支持系统，即智库知识库。

3 基于本体的智库知识库数据集成框架

智库知识库[9]（Knowledge Repository）泛指支持和服务于智库运作的知识库系统，是智库知识能力建设的重要机制。围绕智库研究和服务的决策领域，进行相关知识内容的收集、保存、组织和提供服务，是智库知识库的首要任务；同时，发布和传播智库自身产出的决策咨询产品也是智库知识库的重要功能。因此，智库知识库兼具领域知识库和机构知识库的双重属性和功能——既是智库正常运作及决策产品产出的重要信息支撑工具，也是智库有效管理并利用其知识资产的工具。

实现信息的语义化是大数据下智库数据组织环节的首要目标。通过分析与总结相关文献[10][11]，本文针对智库知识库构建了一个基于决策支持本体的数据集成框架（如图 3），该框架依次按照数据资源→数据集→文档→实体 4 个层次对大数据资源从粒度由粗到细进行描述和组织，描绘了大数据下不同来源不同类型的数据信息经过信息抽取后在智库知识库中被进一步语义化处理，最终都转化为可用于支持决策研究的“智能数据”。智能数据是指通过对海量数据进行处理分析后，从数据中提取出包含有价值的信息和知识，使数据具有“智能”，相比大数据的“大”而言，智能数据拥有更高的数据价值，更值得进行深入挖掘，可通过建立模型寻求现有问题的解决方案或进行预测，“啤酒+尿布”就是一个典型的智能数据应用案例。该数据集成框架的信息处理过程主要包含了以下 5 个阶段：信息抽取阶段、数据存储阶段、数据准备阶段、语义数据模型和应用阶段。

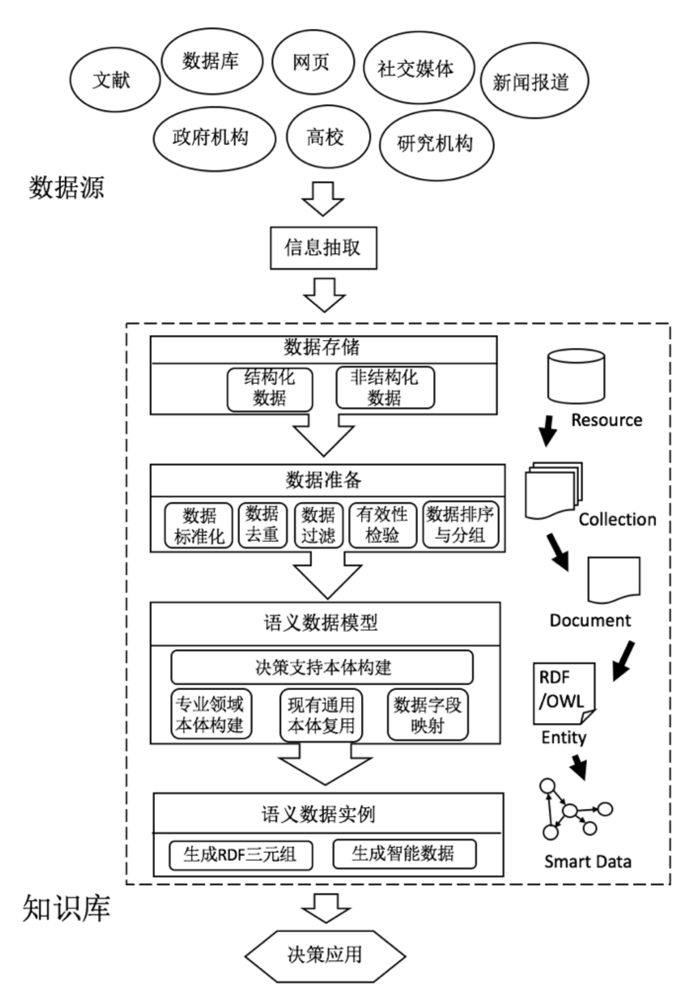


图 3 基于决策支持本体的数据集成框架

Figure 3 Data integration framework based on decision support ontology

3.1 信息抽取阶段

这是智库对大数据下信息资源进行语义化处理过程的第一步，主要是从适当的信息源中抽取相关数据资料，信息源可以是从外界自动抓取数据的外部信息源，也可以是从内部机构人员处收集成果的内部信息源。针对信息资源格式的不同，可以将抽取过程分为结构化抽取和非结构化抽取两部分进行，并分别存储在不同类型的存储库中。

3.2 数据存储阶段

将上一阶段从智库机构内部和外部数据源搜集到的数据进行保存。针对数据的类型分为结构化数据存储和非结构化存储两种方式，存储工具也从传统的数据库管理系统（DBMS）如 MySql、PostgresSQL 等，到企业级数据仓储（EDW）和大规模并行处理数据库（MPP）如 PADB 和 SAND 等，此外 HDFS、HBase 这种分布式文件系统和 MongoDB、CouchDB 等 NoSQL 数据库也经常被用于非结构化数据的存储。

3.3 数据准备阶段

在数据准备阶段,存储设备中的结构化和非结构化信息资源将根据智库机构的服务对象和研究目标按照专题（topic）、学科领域（subject/domain）或项目（project）等被组织成数据集的形式，并在各个数据集中被进一步细分为一个个由文本组成的文档。

这一过程中涉及到对数据的清洗以使数据符合目标模式,其中一些典型的处理方法包括对数据的规范化、数据去重、完整性约束违规检查、基于正则表达式过滤数据、排序和分组数据等等。

3.4 语义模型阶段

语义数据模型是数据集成框架的核心部分,也是面向决策研究的信息资源实现语义化的关键。

在语义数据模型中，支持决策研究的本体将首先被构建，之后依据智库决策研究涉及的具体领域进行专业领域本体的构建,此外还将复用现有的本体和各种通用本体，以针对不断更新的信息对本体模型进行扩展。最后经过本体间数据字段的映射、相似数据字段对齐等一系列流程形成一个面向决策研究过程的通用的语义数据模型。到这一阶段为止,智库知识库的数据组织模型的构建已基本完成,大数据下从任何来源采集到的任何数据都可以经过上述一系列步骤实现面向决策研究的语义化，成为一个个相互关联的实例，从而更有利于决策者挖掘其中的潜在知识，为决策制定提供信息支持。

3.5 决策应用阶段

经过充分语义化后的数据资源已经成为具有较高价值的智能数据,可以根据决策研究过程中的不同需求从不同角度为决策者提供信息支持。与目前传统的较多人工参与的智库决策信息支持机制相比,一方面该信息处理框架利用各种信息处理技术和工具将大数据作为原材料进行深度加工,使其成为能被计算机自动处理的“可计算信息”，逐步实现半自动直至自动化的决策过程。另一方面，得益于大数据巨大的数据体量，以及语义化模块对数据资源的语义化处理，使得更深层次的潜在知识和知识关联得以被挖掘并发现,最终提供给决策者的是智能化的决策支持数据而非一般数据信息,因而这种基于大数据分析的决策研究方法能够得出较传统方法更科学、更可靠也更迅速的决策结果。

4 决策支持本体的构建

本体作为“共享概念模型的明确的形式化规范说明”[12]，其目标是获取、描述和表示相关领域的知识，提供对该领域知识的共同理解，确定领域内共同认可的词汇，并从不同形式的形式化模式上给出了这些词汇（术语）和词汇之间相互

关系的明确定义[13]。针对大数据下智库在信息采集和组织环节面临的海量非结构化数据，本文选择使用本体方法对这些复杂类型数据进行语义化处理，实现本体驱动的决策过程。

面向决策过程的决策支持本体是智库知识库中数据语义模型的核心组成部分，决策支持本体的结构设计包含 3 个阶段：需求分析、本体建模和本体实施。

4.1 需求分析

首先通过文献调研和网络调研识别出围绕决策研究过程的关键问题，对其进行分解，提炼出实体类型和关系类型。在决策研究过程中涉及的实体和关系具有一些重要的特征，对本体的设计提出了相应的要求，主要包括：

（1）决策研究过程中涉及的实体种类较多，如事件、人员、机构、地理位置等，这些实体又可能被进一步细分。例如某次水资源污染事件中，涉及的机构类型可能包括企业、司法部门、工商部门等。此外，事件的关注者有时也是直接参与事件的实体人员或机构，例如当地居民不仅关注并投诉了该事件，也是饮用了污染水的受害者。本体需要具有容纳各种各样相关实体的能力。

（2）与决策研究相关的实体和关系具有时效性，决策议题、相关事件、参与者和关系都存在于特定的时间内。本体需要描述时间维度，以支持对事件发展过程的表示和分析。

（3）对每一个实体和关系，都有大量对应的数据资料为其提供丰富的描述和评论。这些数据资料经过处理整合，可以提供揭示性的定量或定性参考。决策本体有必要将这些基础性的支撑性数据资料也包含在内。

（4）决策本体应保留对其他本体的接口，支持对现有本体和新建本体的扩展。例如科技领域问题会用到学科领域本体，支撑性资料会用到出版物本体。

本体设计的整体要求是在支持上述分析的同时，逻辑模型应尽量简明。

4.2 本体建模

明确需求之后，将要选取合适的构建方法对决策研究问题进行本体建模。面向决策研究过程的本体设计就是根据决策支持本体的构建目标建立其概念模型的过程。构建决策支持本体的目标是建立基于决策支持信息的语义检索系统，为智库专家和政策研究人员提供语义化的信息查询方式，突破传统的智库决策信息支持机制，提供语义级的决策信息查询服务。

本文根据决策研究问题的实际情况和需要，选择国际上较为成熟的七步法作为参照主体来构建科技智库知识库中的决策支持本体。具体构建步骤如下：

（1）确定本体的范畴和目的

界定决策支持本体的范畴，即要明确如何描述一个决策相关的事件或资源，以及描述到何种程度。目的是希望能够建立一个通用的面向决策研究过程的本体，

用以描述针对某决策议题或事件引发的问题，以及针对问题作出的回应、涉及到的项目、相关参与人、机构及其相互之间的关系等等。

（2）考虑复用现有本体的可能性

本文选择复用 DC 和 ABC 本体。都柏林核心 DC 作为目前使用最为广泛的本体之一，其对资源基本情况的语义描述具有很广泛的适用性和扩展性。改变模型的能力使得 ABC 本体适合于描述各种各样的实体和它们之间的关系，包括所有媒体类型的对象（文本、图像、视频、音频、网页和多媒体等）。它还可以用于模拟诸如知识内容和时间实体的抽象概念，例如对象发生的性能或生命周期事件。因此本文最终选择复用 DC 和 ABC 本体中的相关类和属性，同时利用 XML schema、RDF schema、OWL 等命名空间。所有复用的本体如表 4 所示：

表 4 复用本体

Table 4 Reused ontologies

命名	值
xmlns:dc	“http:// http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements”
xmlns:abc	“http://dcpapers.dublincore.org/pubs/article/view/655/651”
xmlns:rdf	“https://www.w3.org/TR/2014/REC-rdf-syntax-grammar-20140225/”
xmlns:owl	“https://www.w3.org/2002/07/owl#”
xmlns:xsd	“https://www.w3.org/2001/XMLSchema#”
xmlns:rdfs	“https://www.w3.org/TR/2014/REC-rdf-schema-20140225/”
xmlns:protege	“http://protege.stanford.edu/plugins/owl/protege#”

（3）列出本体中的重要术语

确定本体中核心概念的具体表述词汇及其逻辑关系，最常见的方式就是直接抽取对应领域主题词表和分类表中的主题词和分类词。

（4）列出关键实体和类

对提取出的核心概念进行评估，按照一定的逻辑规则进行分组，设计合理的类及其层次结构。本文通过参考相关研究的论文及研究结果[14][15]，结合决策研究过程的实际特征，最后确定从问题类（Issue）、决策产品类（Decision output）、决策建议类（Decision suggestion）、参与者类（Participant）、资源类（Resource）这 5 个核心方面构建决策本体，并据此展开整个类层次结构。

问题类（Issue）是指由科技领域或其他相关领域事件（Event）引发产生的各种问题。

决策产品类（Decision output）是指智库研究人员及决策者针对产生的问题进行科技政策研究，最终得到的决策产品以及在研究过程中产生的各种中间数据。具体分类体系如图 4 所示。

决策建议类（Decision suggestion）是科技智库决策研究过程的最终产物，是决策产品类的一个子类。决策产品的另一个子类是中间产品类(Mid-product)，指在决策者进行政策制订的过程中原始数据经参与研究的智库专家及研究人员的加工生成一系列为其提供思路的中间数据，这些中间数据对于今后类似项目的研究有很大参考价值，通常也被智库进行组织并保存。

参与者类（Participant）是指所有直接或间接参与到决策研究过程中的个人（Individual）或组织（Organization），其分类体系如图 5 所示。

资源类（Resource）是指在决策研究过程中为决策产品的生成提供支持的各类信息资源，包括各种数据（data）、方法（method）、模型（model）、工具（tool）等，同时也为科技问题的溯源和询证（evidence-based）提供了途径，详细分类体系如图 6 所示。

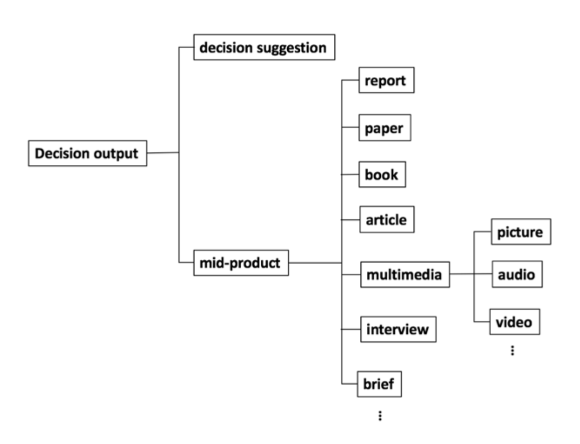


图 4 决策产品分类

Figure 4 The classification of decision output

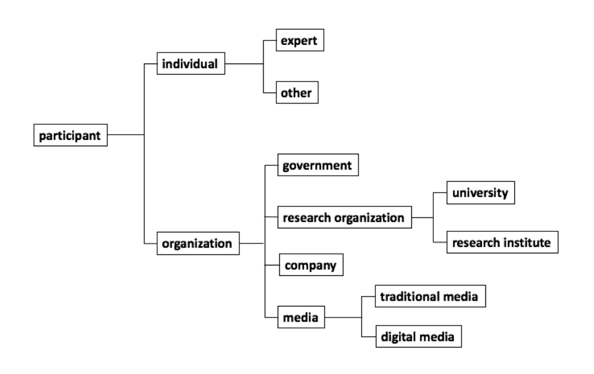


图 5 参与者分类

Figure 5 The classification of participants

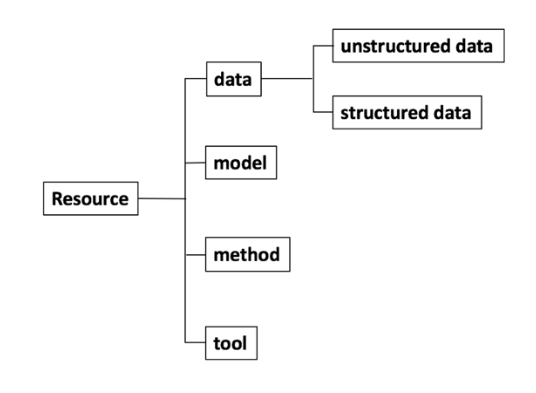


图 6 资源分类

Figure 6 The classification of resources

最终构建的决策支持本体的总体框架见图 7，整个本体框架分成三层：核心层、扩展层和支撑层：

①核心层——问题 Issue、决策产品 Decision output，

②扩展层——参与者 Participant、事件 Event、项目 Project、任务 Task、决策建议 Decision suggestion、中间产品 Mid-product 等，

③支撑层——资源 Resource、数据 Date、模型 Model、方法 Method、工具 Tool 等。

分层结构提供了简明的逻辑模型，使得核心层、扩展层和支撑层的实体关系清晰有序；不同层次存储的数据结构复杂度和精确度不同，允许系统根据查询需求对准确性和全面性的权衡，满足个性化的查询和分析结果。

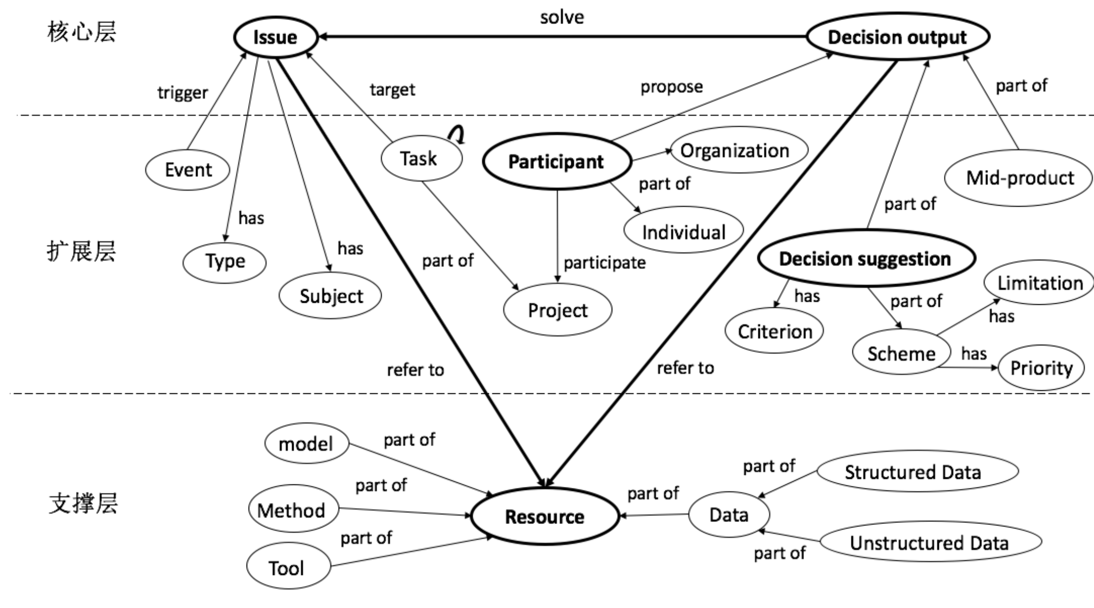


图 7 科技智库的决策支持本体总体框架

Figure 7 Decision-making ontology of science and technology think tanks

（5）分析实体的属性

本体的数据属性是表示类或概念与值的关系，例如年龄属性“at age of”的数值将代表某种生物的具体年龄，而本体的对象属性则表示类之间的非等级关系，例如属性“trigger”或“cause”可以用以表示两个类之间的因果关系，一个类触发了另一个类，恰当地定义数据属性和对象属性可以有效地反映类间的关系。限于篇幅仅展示核心层的部分属性如下表 5：

表 5 核心层类的部分属性

Table 5 Parts of properties of classes in the core layer

类名	数据属性
问题类 Issue	编号（issue_id），名称（issue_name），类型（issue_type），主题（issue_subject），内容（issue_content）
决策产品类 Decision output	编号（output_id），名称（output_name）
类名	对象属性
解决 solve	决策产品编号（output_id），问题编号（issue_id）

（6）分析属性的约束

对属性进行必要的约束限制，针对数据属性的约束条件包括描述属性的值的类型（字符串、布尔型、枚举型等）、值域、基数（单个基数或多个基数）等特性。例如将时间属性的 Year 字段的最大值设为 2017，又比如人的性别只能从“男”或“女”两个值中选择一项等。

（7）创建实例

根据之前步骤已经建立的概念模型创建具体的实例。本文使用 protégé4.3 选择一个具体领域的问题进行部分实例添加作为展示。此外，也可借助 API 工具实现对实例的批量导入。本文将创建的本体以 OWL 文本的格式保存在本地计算机中以便于本体的复制与备份，并可随时进行编辑和修改。

4.3 智库知识库的实施过程

智库知识库的总体运作流程见图 8（以科技智库为例）。通过 API 工具和爬虫工具可以在互联网上抓取所研究领域的相关事件和问题内容，同时收集用户的政策需求或政策方案，对自然语言进行处理后，再通过信息抽取等文本处理工具分析上述内容，提取其中的实体和类型，使用上述本体构建模型和 Protégé等编辑工具进行本体的构建与数据维护，并且可依据信息源的性质及数据特征引入外

部本体，如在本例中的 DC 元数据和科技领域本体。最终将从现实世界所采集的内容全部实现实体化，形成相互具有语义关联的智能数据，结合之前用户的政策需求或政策方案提供相应的匹配结果，为政策制定者提供决策支持信息。最终预期达到的目标是能够基于智库采集处理的大数据信息构建一套面向决策研究过程的语义驱动的检索系统（图 9）。

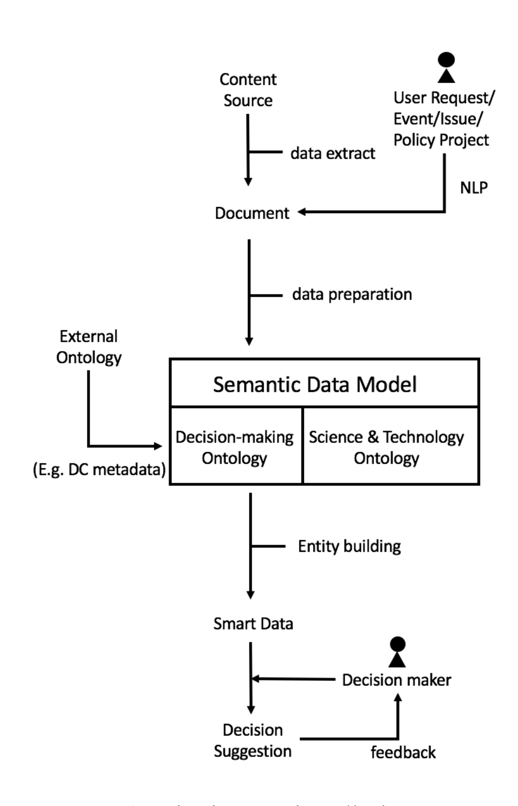


图 8 智库知识库运作流程

Figure 8 Knowledge repository operation process of think tanks

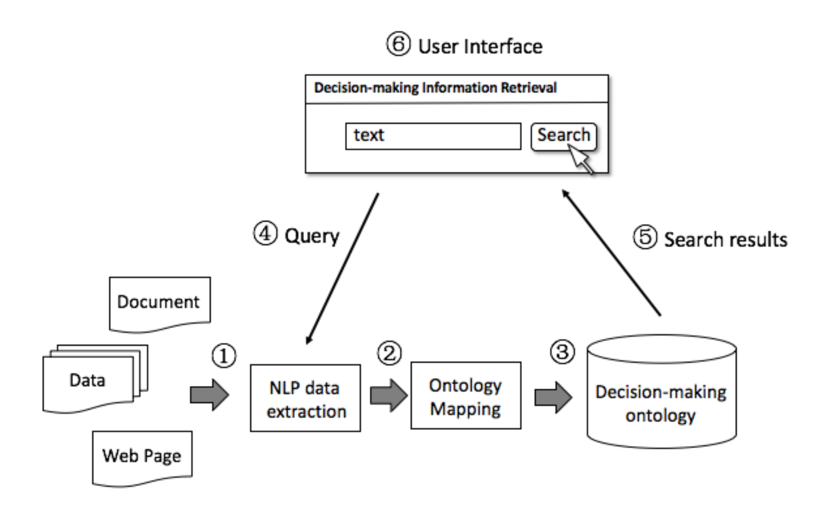


图 9 语义驱动的决策检索系统示例

Figure 9 An example of a semantic-driven decision-making retrieval system

在这套检索系统中，社会热点事件、某个社会问题、新闻报道、研究出版物等大数据下的海量数据资料被智库抽取并进行自然语言处理，再经过智库知识库的处理并存储，成为智能数据。在检索时，用户的查询需求同样先进行自然语言处理，之后请求数据以相同的方式被映射到智库知识库中的决策本体中，继而匹配之前生成的实体化数据，返回给用户基于智能数据的检索结果，这种基于语义驱动和智能数据的检索系统相比传统智库在决策研究中的资料查找和收集方式，能够得到更为全面、更为科学、相互关联的检索结果，概括来讲就是基于上下文情境（context based）的结果，这种检索结果可以更容易地进行潜在知识挖掘，带有情境之后也更方便被直接投入决策应用。例如，对“雾霾防治”相关信息进行检索，则通过该语义驱动的决策检索系统，用户将会得到某次雾霾发生事件的起因、起止时间、地点、相关人员、相关机构、相关出版物描述，以及雾霾防治政策的相关研究机构、研究人员、政策实施情况、对时间产生的影响等等一系列信息。由此实现了一种面向决策研究的、提供上下文情境的语义化的检索机制。打破了传统的较依赖人工的智库决策信息支持机制。

5 总结

本文主要调研了当前智库传统信息支持机制及其在大数据时代背景下受到的挑战，从而提出了一种一般性的语义化的智库决策信息处理框架，并依据该框架设计了科技智库知识库的数据组织模式，以期为大数据下新型智库的建设提出参考借鉴。

关于针对智库政策领域研究的本体创建，目前在国内的尝试还少之又少，后续要进行的工作还有很多。在不断完善该智库决策信息处理框架功能并维护决策本体的同时，本文下一步的研究工作将围绕决策本体的本体映射、本体评价、集成扩展等方面继续进行。

参考文献：

[1]James G. McGann, University of Pennsylvania, 2016 Global Go To Think Tank Index Report[EB/OL].[2016-08-10]. http://repository.upenn.edu/think_tanks/10/.

[2]吴育良.国外智库决策信息支持研究及启示[J]. 图书馆理论与实践,2015(10):31-35.

[3]廖球,严扬帆,莫崇菊.大数据时代机构自建学术数据库研究[J].图书馆学刊,2014(4):34-36.

[4]MOORTHY J, et al. Big data: prospects and challenges[J]. The Journal for Decision Makers, 2015,40(1):74-96.

[5]Bruegel. Datasets[EB/OL]. [2017-03-01].http://bruegel.org/publications/datasets/.
[6]SIPRI. Databases[EB/OL]. [2017-03-01].https://www.sipri.org/databases.
[7]许鑫,吴珊燕.智库知识库的构建研究[J].情报理论与实践,2014(3):68-72.
[8]Studer R, Benjamins C R, Fensel D. Knowledge Engineering, Principles and Methods[J]. Data and Knowledge Engineering,1998, 25(1-2):161-197.
[9]Benjamin Horne, Tanya Torres, Jessica Mackenzie. Think Tank Management: Establishing a Knowledge Repository[EB/OL].[2016-12-30]. http://www.ksi-indonesia.org/en/news/detail/think-tank-management-establishing-a-knowledge-repository.
[10]Srividya K Bansal, Sebastian Kagemann. Semantic Extract-Transform-Load framework for Big Data Integration[J]. Computer, 2015,48(3):42-50.
[11]Francesco Corcoglioniti, Marco Rospocher, etc. The KnowledgeStore: A Storage Framework for Interlinking Unstructured and Structured Knowledge[J]. International Journal on Semantic Web and Information Systems,2015,11(2):1-35.
[12]Studer R, Benjamins C R, Fensel D. Knowledge Engineering, Principles and Methods[J]. Data and Knowledge Engineering. 1998, 25(1-2)
[13]廖军.基于领域本体的信息检索研究[D].长沙:中南大学,2007.
[14]Leo Wanner, Marco Rospocher, etc. Ontology-centered environmental information delivery for personalized decision support[J]. Expert System with Applications.2015(42):5032-5046.
[15]Studer R, Benjamins C R, Fensel D. Knowledge Engineering, Principles and Methods[J]. Data and Knowledge Engineering. 1998, 25(1-2).

The Challenges and Data Organization Strategies for Foreign Think Tanks within the Context of Big Data

AnNan[1,2] [1]Zhu Zhongming[1]
[1] Lanzhou Library, National Science Library of Chinese Academy of Sciences, Lanzhou 730000
[2] University of Chinese Academy of Sciences, Beijing 100049

Abstract: [Purpose/significance] The high level of new types of think tanks cannot be separated from the high level information support mechanism. The traditional think tank information organization mechanism in the big data era cannot adapt to the current data characteristics and decision-making requirements. Constructing the knowledge repository that supports the decision-making process is becoming the inevitable trend of the development of think tanks. **[Method/process]** This paper chose foreign think tanks with reference value in *2015 Global Go To Think Tank Index Report* as the research objective, and summarized several kinds of information organization methods which were common in the current think tank by using the literature research method and case analysis method. It also analyzed the data value chain and its requirements for all aspects of the think tank, and accordingly put forward the necessity of the construction of knowledge repository of think tanks. **[Result/conclusion]** Finally, a general knowledge repository framework for decision-making process is proposed, and the knowledge organization model in the knowledge repository is constructed by the semantic ontology method. In order to provide references for the achievement of the transformation from the semi-automatic decision-making process to the automatic one.

Keywords: think tank knowledge repository big data information organization organizational strategy decision research ontology